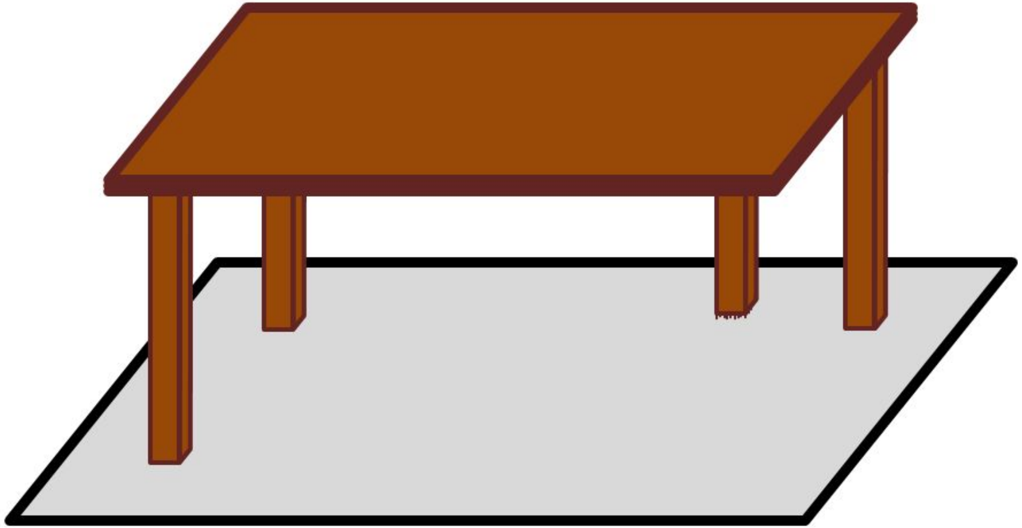
A new paradigm for computer vision based on compositional representation

Vinjai Vale
Mentor: Kevin Ellis

Computer vision today

- Image processing, analysis, understanding
- State of the art: deep convolutional neural network (CNN)
- CNNs excel at *classification*, struggle with *representation*

Example



Object compositionality

Represent objects recursively through their components and relations

Enables complex human-level visual reasoning

CNN drawbacks

CNNs learn fuzzy patterns on textures — poor spatial understanding

https://rocknrollnerd.github.io/ml/2015/05/27/leopard-sofa.html
https://sleezybarbhorsewear.com/customers-page/

T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. CoRR, abs/1606.03498, 2016.

A dog is jumping to catch a frisbee.

A refrigerator filled with lots of food and drinks.

A yellow school bus parked in a parking lot.

Research goal

- Engineer domain and dataset to isolate the problem of compositionality

- Develop algorithms, techniques to solve representation tasks on dataset

Primitive elements and ShapeWorld

A small number of *primitive elements* (PEs) compose all objects.

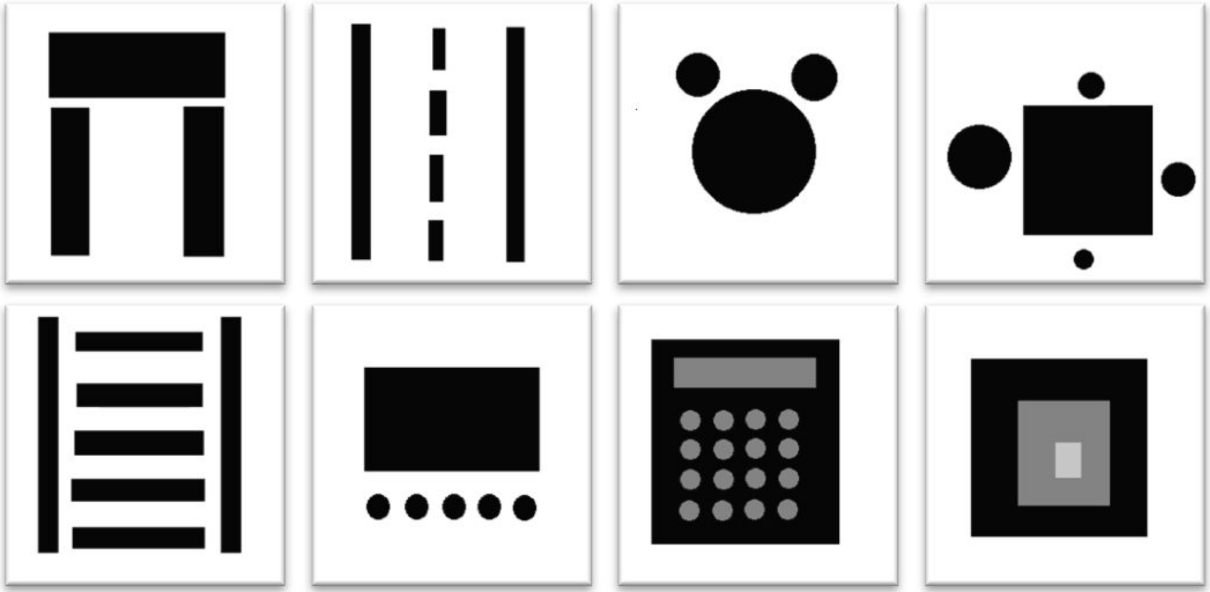
Biederman 1987: 36 *geons* are the PEs of human vision in 3D world

Key intuition:

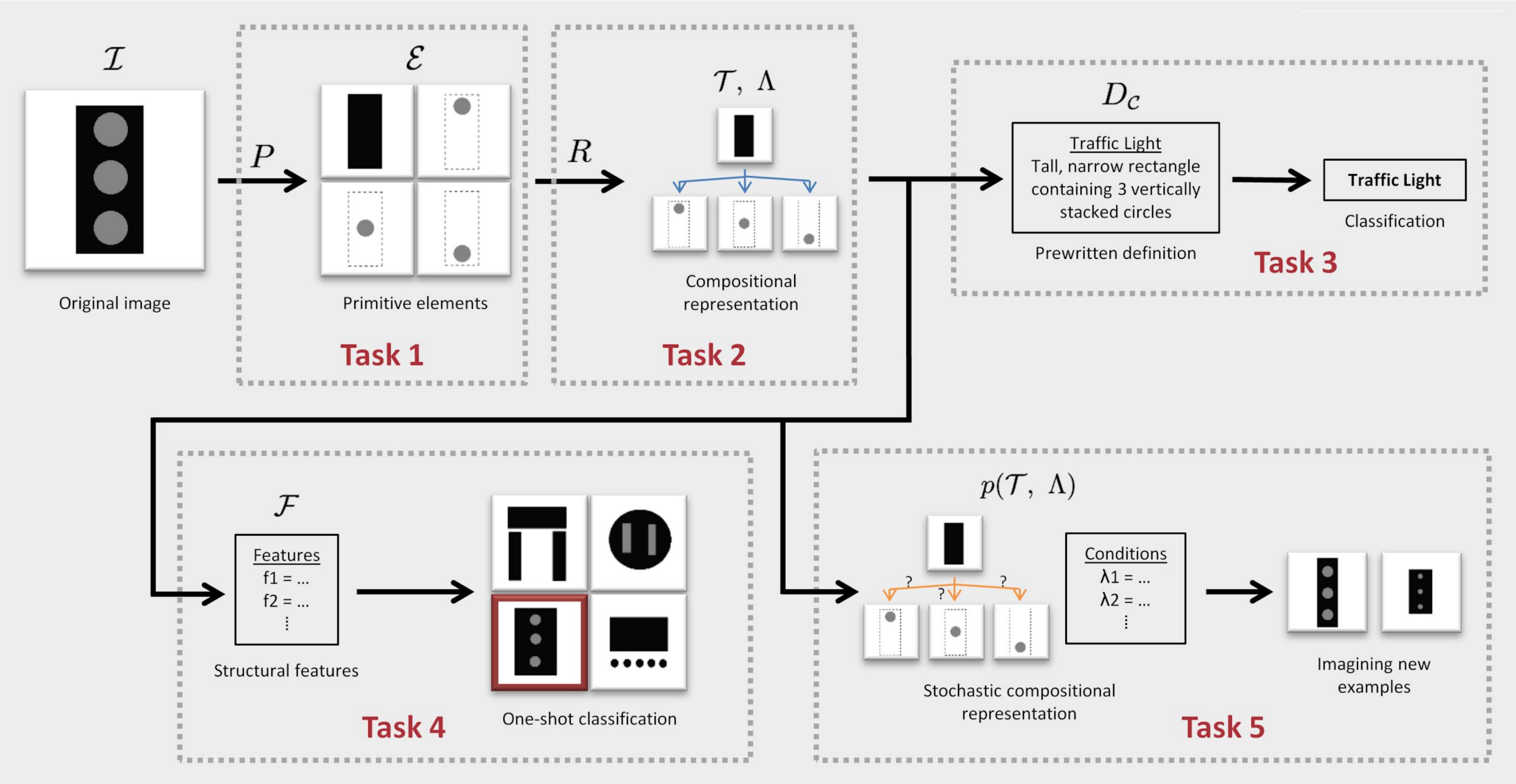
- **3D world with 36 PEs is hard...**
- **Instead, solve vision in 2D world with only 2 PEs!**

Primitive elements and ShapeWorld

ShapeWorld: 2D dataset composed of circles and rectangles

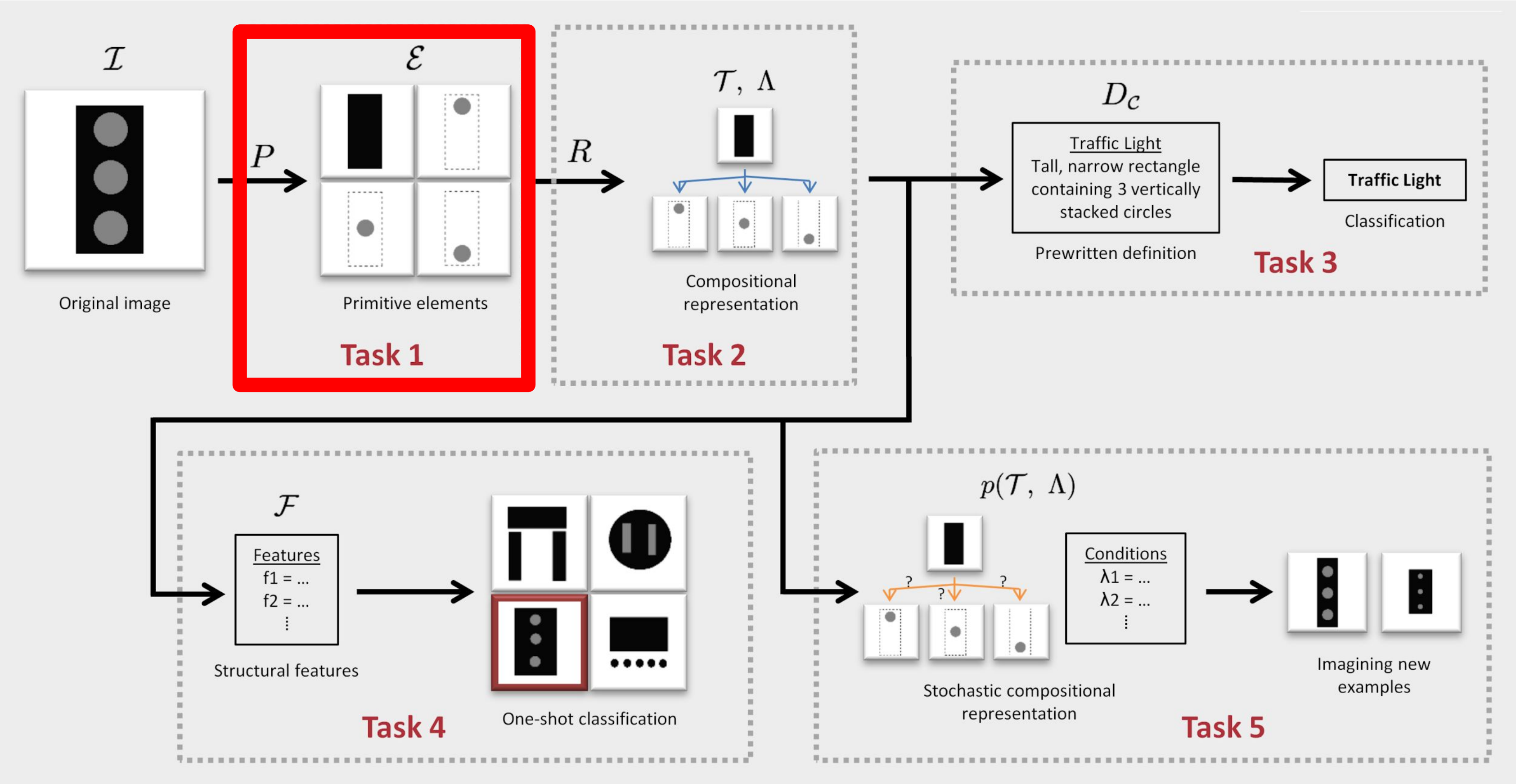


Five Tasks



$\mathcal{I}$
Original image
P
$\mathcal{E}$
Primitive elements
Task 1
R
$\mathcal{T}, \Lambda$
Compositional representation
Task 2
D_C
Traffic Light
Tall, narrow rectangle containing 3 vertically stacked circles
Prewritten definition
Traffic Light
Classification
Task 3
$\mathcal{F}$
Features
f1 = ...
f2 = ...
⋮
Structural features
One-shot classification
Task 4
$p(\mathcal{T}, \Lambda)$
?
?
?
Stochastic compositional representation
Conditions
λ1 = ...
λ2 = ...
⋮
Imagining new examples
Task 5

Task 1



$\mathcal{I}$
Original image
P
$\mathcal{E}$
Primitive elements
Task 1
R
$\mathcal{T}, \Lambda$
Compositional representation
Task 2
D_C
Traffic Light
Tall, narrow rectangle containing 3 vertically stacked circles
Prewritten definition
Traffic Light
Classification
Task 3
$\mathcal{F}$
Features
f1 = ...
f2 = ...
Structural features
One-shot classification
Task 4
$p(\mathcal{T}, \Lambda)$
Conditions
λ1 = ...
λ2 = ...
Stochastic compositional representation
Imagining new examples
Task 5

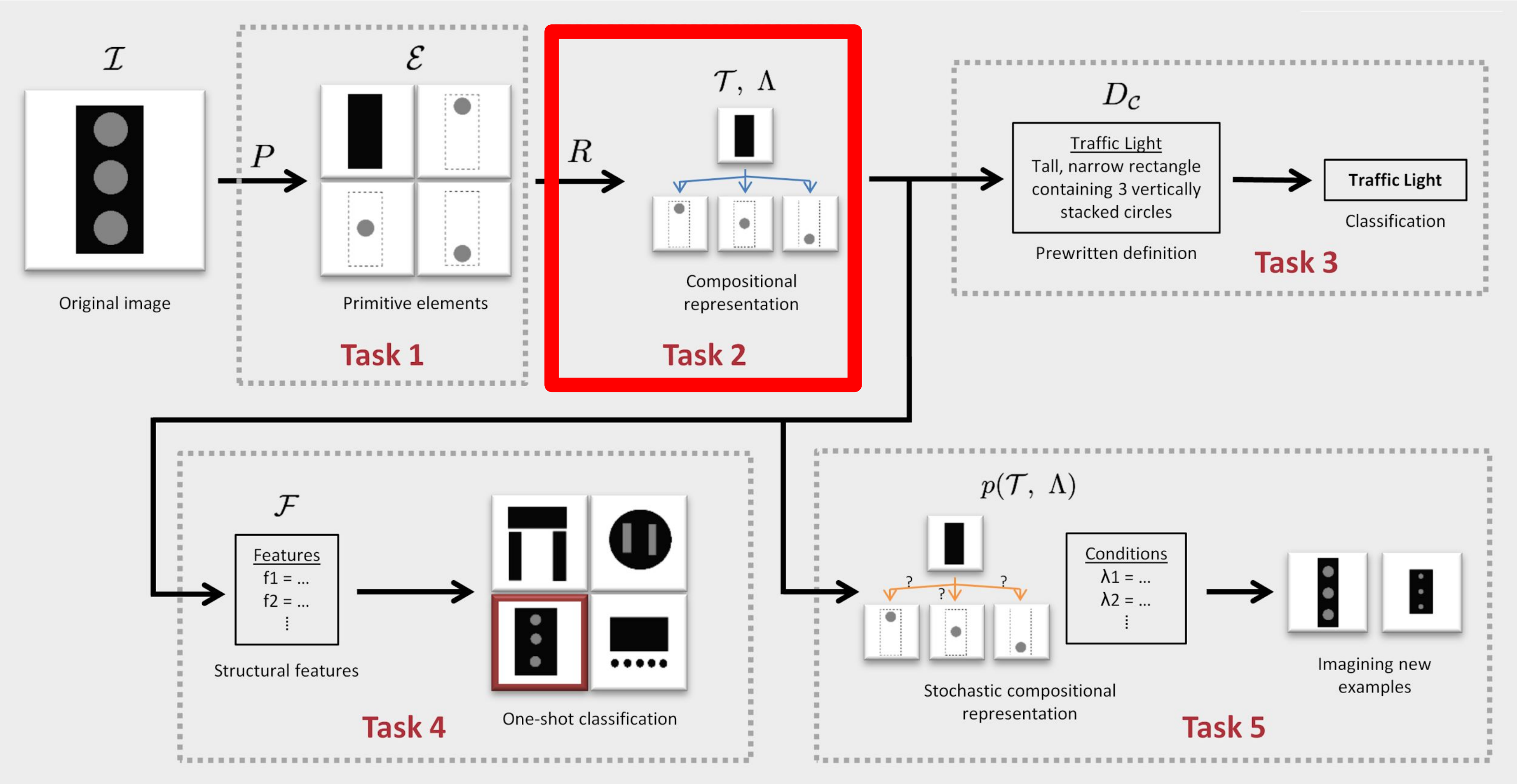
Task 1: PE decomposition

Represent an image in terms of a *graphics program*

“Draw circle here with these coordinates,” “draw rectangle there with those coordinates,” etc.

Image processing methods - OpenCV

Task 2

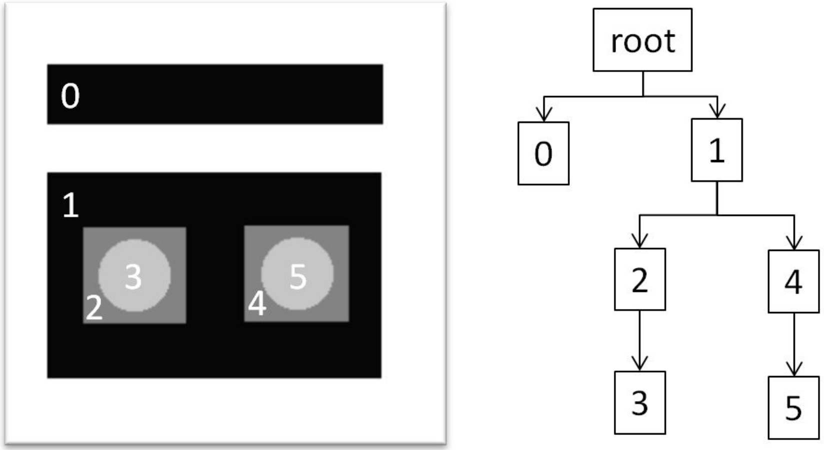


$\mathcal{I}$
Original image
P
$\mathcal{E}$
Primitive elements
Task 1
R
$\mathcal{T}, \Lambda$
Compositional representation
Task 2
D_c
Traffic Light
Tall, narrow rectangle containing 3 vertically stacked circles
Prewritten definition
Traffic Light
Classification
Task 3
$\mathcal{F}$
Features
f1 = ...
f2 = ...
Structural features
One-shot classification
Task 4
$p(\mathcal{T}, \Lambda)$
Conditions
λ1 = ...
λ2 = ...
Stochastic compositional representation
Imagining new examples
Task 5

Task 2: Compositional representation

Two-part data structure to encompass compositional relationships

Augmented Primitive Element Tree (APET)

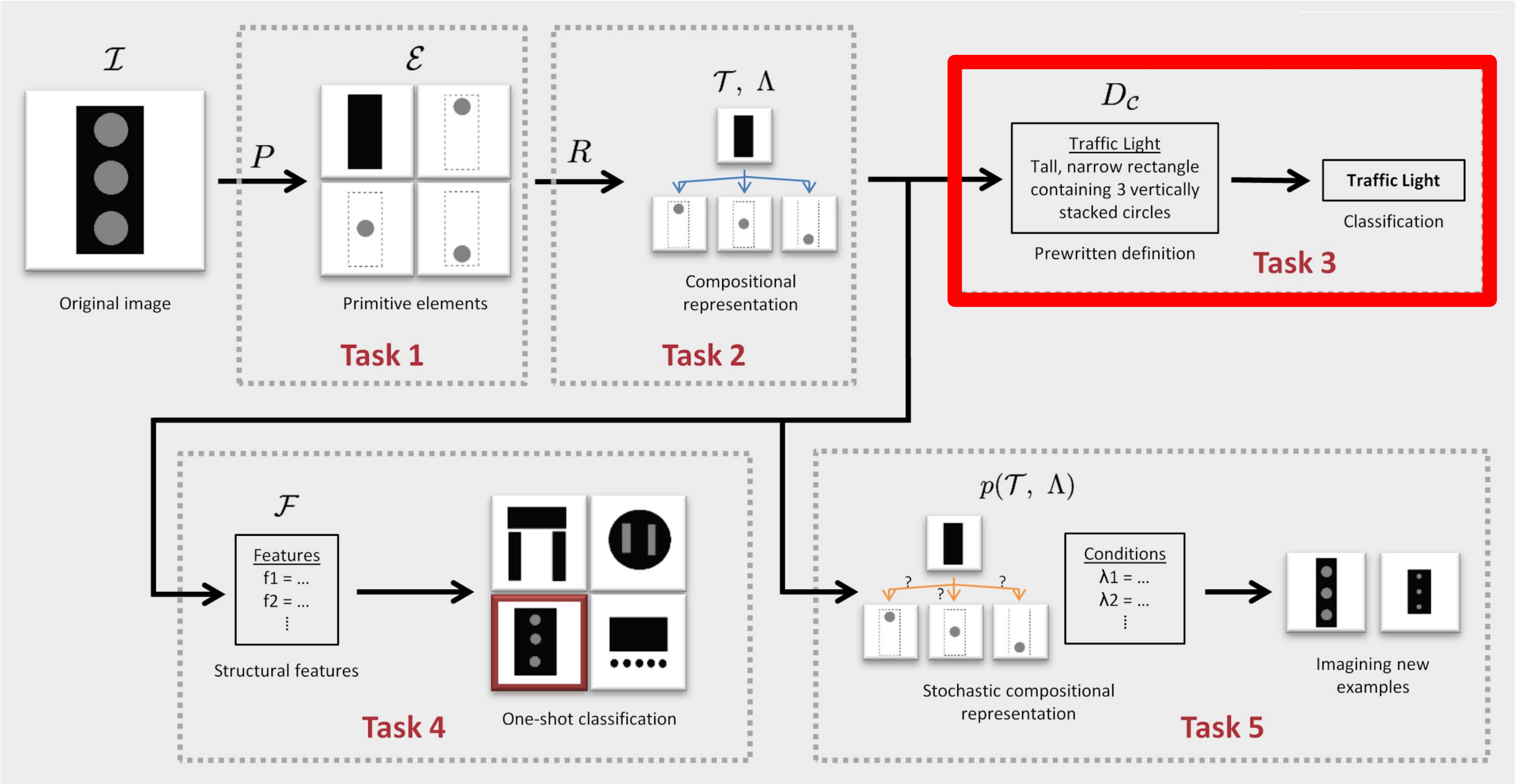


Coincidence List

Coincidences likely part of underlying concept

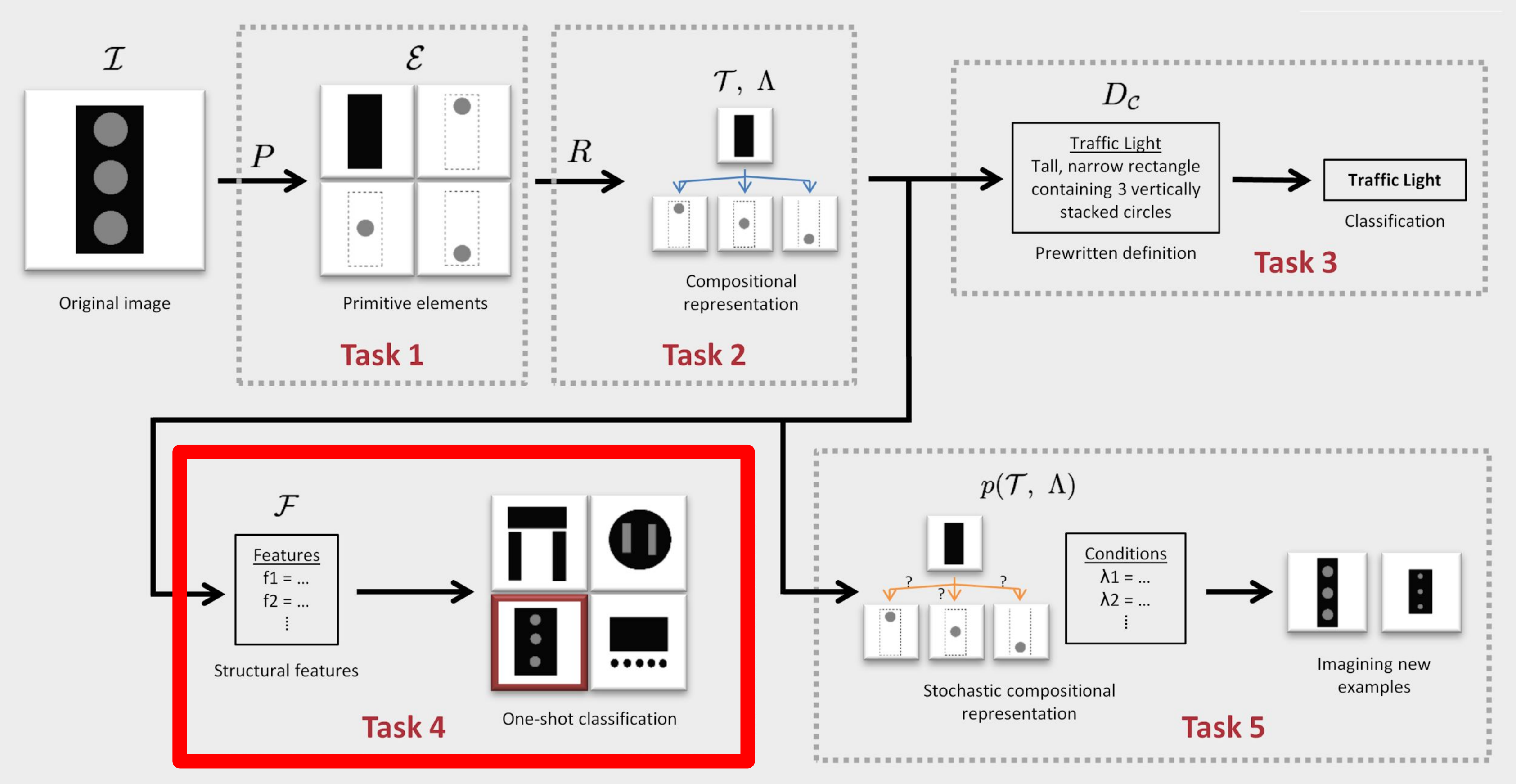
- Rows
- Clusters
- Grids
- Radial arrangement
- Etc.

Task 3



$\mathcal{I}$
Original image
P
$\mathcal{E}$
Primitive elements
Task 1
R
$\mathcal{T}, \Lambda$
Compositional representation
Task 2
D_C
Traffic Light
Tall, narrow rectangle containing 3 vertically stacked circles
Prewritten definition
Traffic Light
Classification
Task 3
$\mathcal{F}$
Features
f1 = ...
f2 = ...
Structural features
One-shot classification
Task 4
$p(\mathcal{T}, \Lambda)$
?
?
?
Conditions
λ1 = ...
λ2 = ...
Stochastic compositional representation
Imagining new examples
Task 5

Task 4

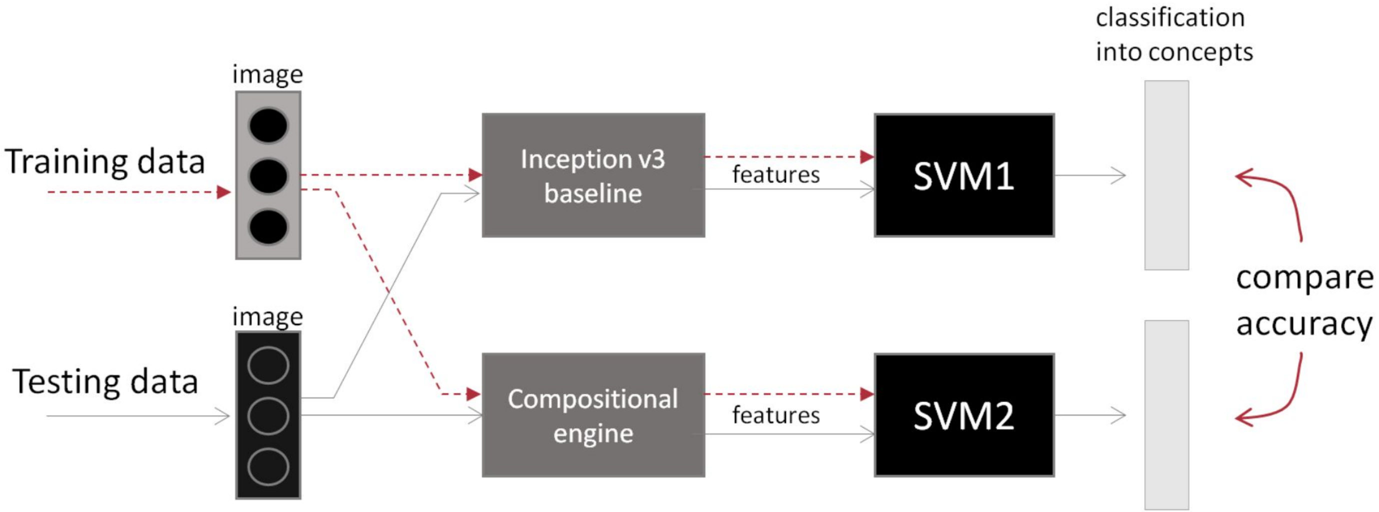


$\mathcal{I}$
Original image
P
$\mathcal{E}$
Primitive elements
Task 1
R
$\mathcal{T}, \Lambda$
Compositional representation
Task 2
D_C
Traffic Light
Tall, narrow rectangle containing 3 vertically stacked circles
Prewritten definition
Traffic Light
Classification
Task 3
$\mathcal{F}$
Features
f1 = ...
f2 = ...
Structural features
Task 4
One-shot classification
$p(\mathcal{T}, \Lambda)$
?
?
?
Conditions
λ1 = ...
λ2 = ...
Stochastic compositional representation
Imagining new examples
Task 5

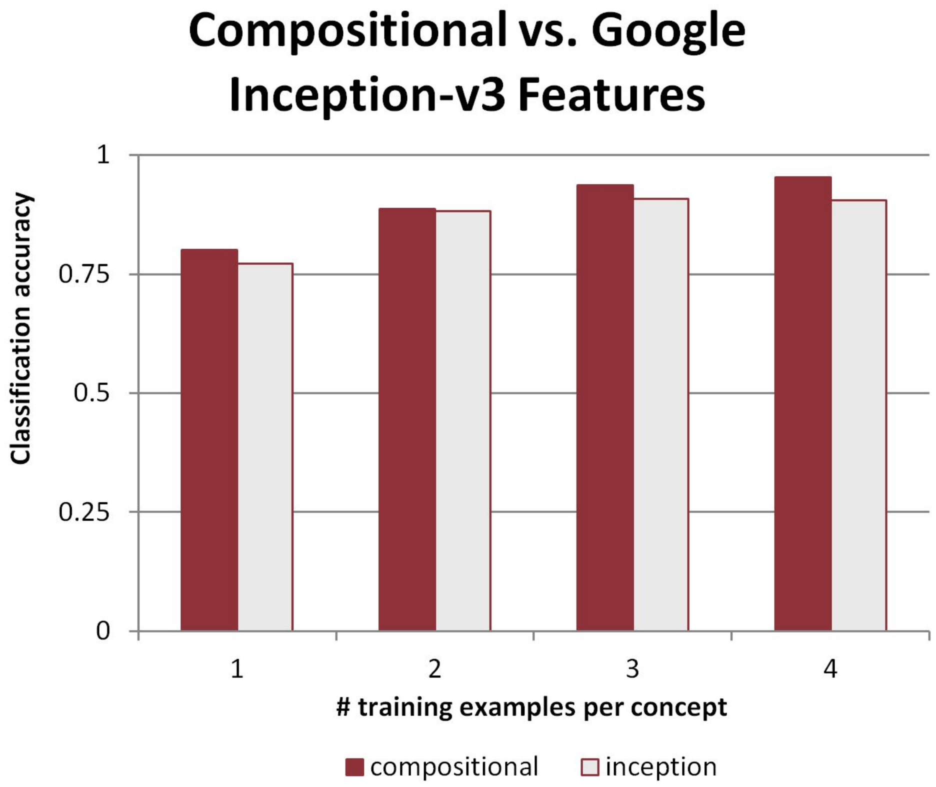
Task 4: Baseline comparison

Extracted 34 features from Task 2 compositional representation, compared against 2048 features from Google Inception-v3 CNN

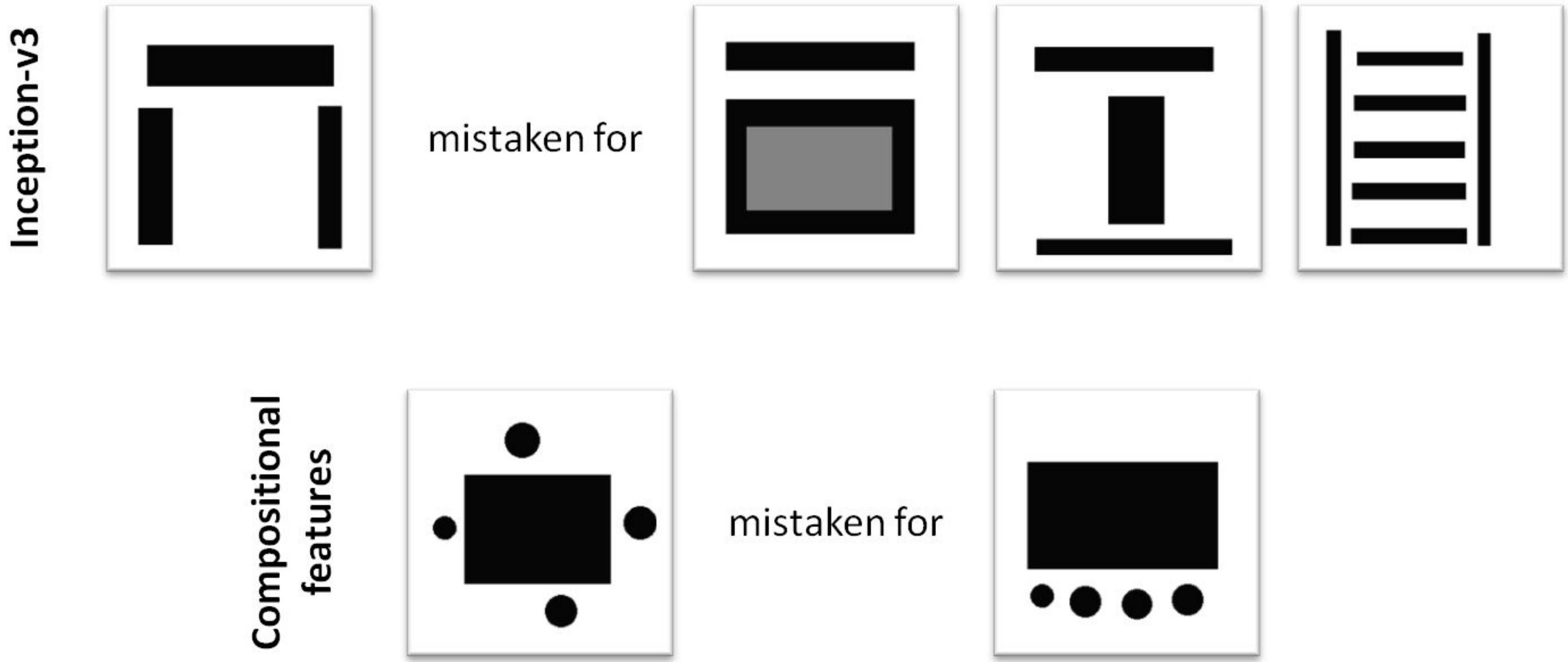
The task: learn to classify a ShapeWorld concept only given 1-4 examples



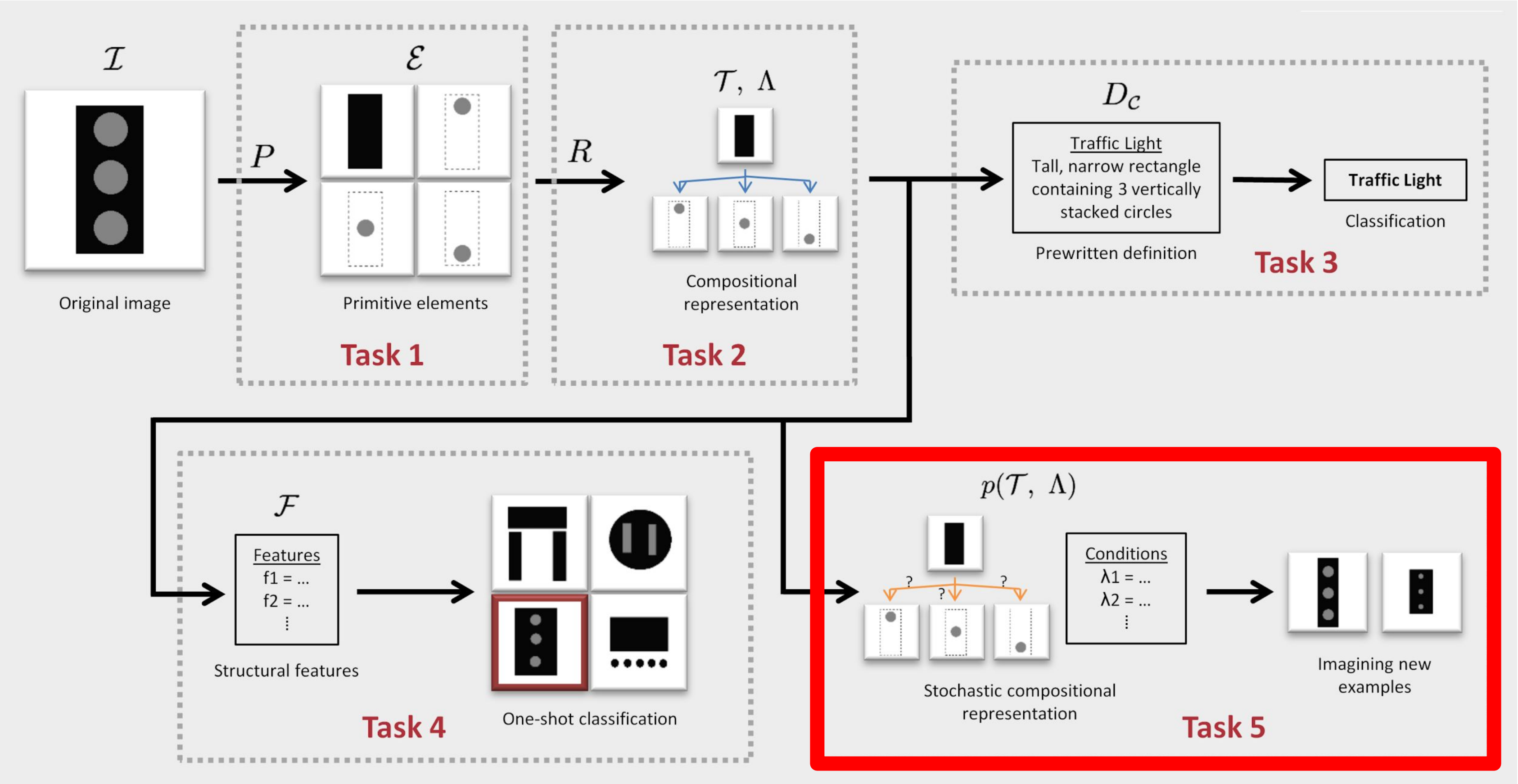
Task 4 results



Task 4 closer inspection



Task 5

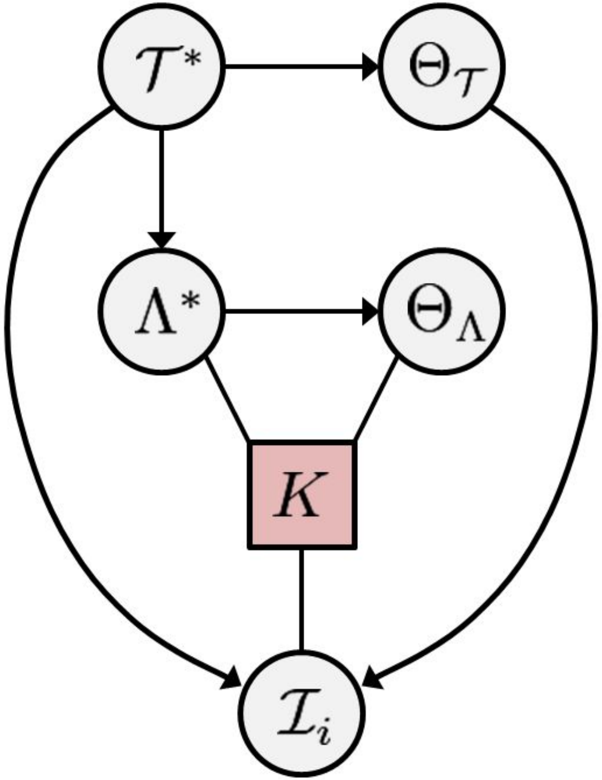


$\mathcal{I}$
Original image
P
$\mathcal{E}$
Primitive elements
Task 1
R
$\mathcal{T}, \Lambda$
Compositional representation
Task 2
D_C
Traffic Light
Tall, narrow rectangle containing 3 vertically stacked circles
Prewritten definition
Traffic Light
Classification
Task 3
$\mathcal{F}$
Features
f1 = ...
f2 = ...
Structural features
One-shot classification
Task 4
$p(\mathcal{T}, \Lambda)$
Conditions
λ1 = ...
λ2 = ...
Stochastic compositional representation
Imagining new examples
Task 5

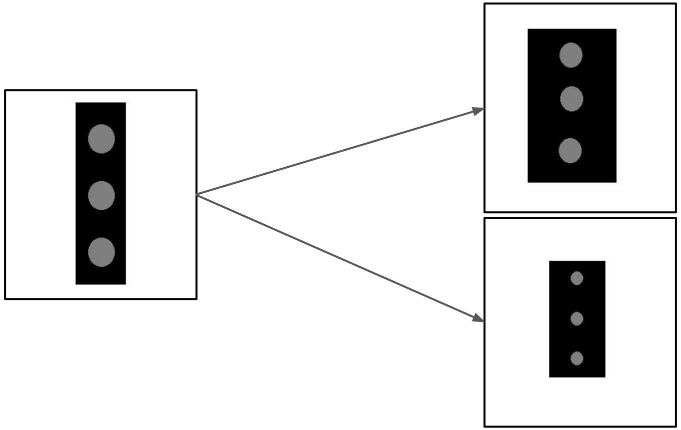
Task 5: Learning a generative model

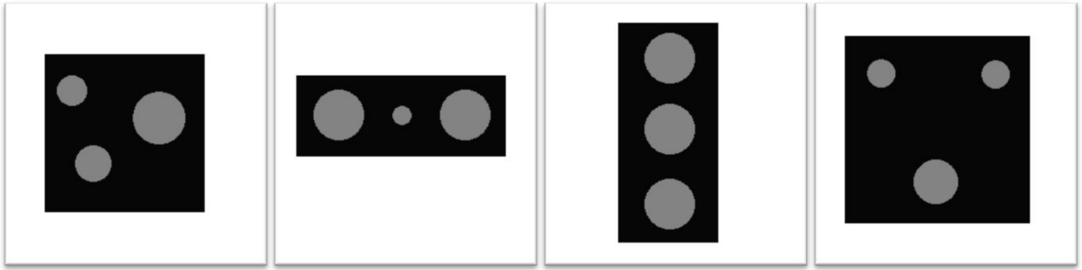
Mathematical model for inferring a probability distribution $P(I, S_c)$: probability that image I belongs to concept C

Task 5: Learning a generative model

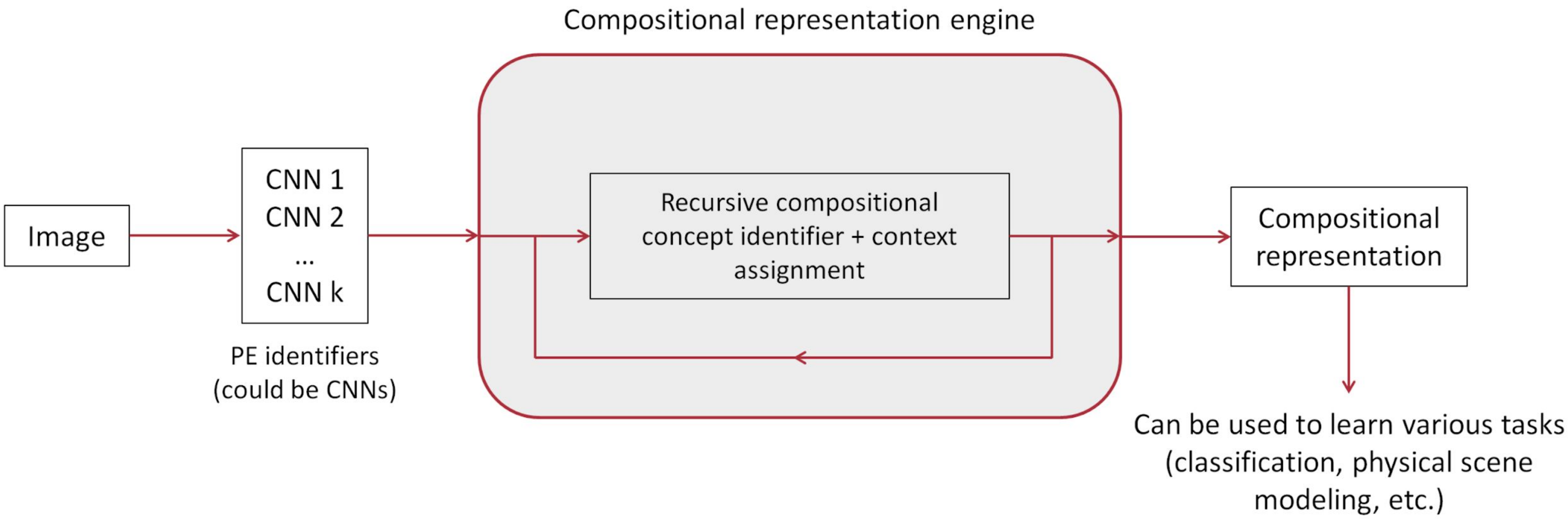


Task 5: Learning a generative model





Beyond



Compositional representation engine
Image
CNN 1
CNN 2
...
CNN k
PE identifiers
(could be CNNs)
Recursive compositional concept identifier + context assignment
Compositional representation
Can be used to learn various tasks (classification, physical scene modeling, etc.)

Conclusion

1. ShapeWorld dataset following Recognition-by-Components paradigm

2. Generate compositional representations of images (APET, Coincidence List)

3. Compositional representation feature set outperforms state-of-the-art CNN in object compositionality tasks

4. Mathematical, generalizable proabilistic approach to learning stochastic compositional representations

Goal: AI vision systems that are faster, safer, and closer to human vision.

Acknowledgements

Mentor Kevin Ellis @ MIT Brain and Cognitive Sciences

MIT PRIMES Program

Szczesny Kaminski and Sean Campbell, my teachers, for facilitating independent studies in machine learning